ISSN 1561-5359. Штучний інтелект, 2016, № 4

УДК 004.93

О.Г. Марголін
Київський національний університет імені Тараса Шевченка, Україна
пр. Академіка Глушкова, 4д, м. Київ, 83000

СИСТЕМА ВИЯВЛЕННЯ ІНФОРМАЦІЇ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ КОРИСТУВАЧІВ

A.G. Margolin
Taras Shevchenko National University of Kyiv, Ukraine
4d, Academician Hlushkov av., Kyiv, 83000

INFORMATION ANALYSIS SYSTEM FOR USER TEXT MESSAGE

Подані результати одного з підходів реалізації ідентифікації та аутентифікації суб’єктів автоматизованої системи через аналіз текстової інформації, що вводить користувач під час листування, коментування, та написання статей. Для вирішення цієї проблеми запропонована реалізація системи аналізу текстової інформації та прийняття рішень. Система може бути використана на практиці, наприклад для виявлення інтернет «ботів» в соціальних мережах, формах та порталах новин.

Ключові слова: ідентифікація користувача, аналіз текстової інформації, автоматизована система.

This article is devoted to research of methods and tools for identifying subjects of an automated system, through the analysis of textual information, that the user enters in the correspondence, commentary, and writing articles. This problem remains unresolved, because there is no single method for identifying a person who is using the computer at the moment. The system can be used in practice, for example to identify Internet “bots” in social networks, forums or news portals.

Keywords: user identification, information environment, decision-making system, analysis of textual information, automated system.

Вступ

Одна з основних проблем, що постає перед розробниками програмних додатків - ідентифікація користувача. Реєстрація користувача полегшує збір статистики і забезпечує цілісність даних користувача, що дає можливість власникам додатків вигідно для себе і для користувача використовувати цю інформацію. Ідентифікацію та аутентифікацію можна вважати основою програмно-технічних засобів безпеки, оскільки інші сервіси розраховані на обслуговування іменованих суб'єктів [1]. Ідентифікація та аутентифікація - це «перша лінія оборони» інформаційного простору організації.

Ідентифікація дозволяє суб'єкту (користувачеві, процесу, що діє від імені користувача, чи іншого апаратно-програмному компоненту) назвати себе (повідомити своє ім'я). За допомогою аутентифікації друга сторона переконується, що суб'єкт дійсно той, за кого він себе видає. Як синонім слова «аутентифікація» іноді використовують словосполучення «перевірка справжності» [2].

Попередня апробація вже досягнутих результатів показала, що одним з найвагомішим фактором ідентифікації (якщо брати за приклад сайти з можливістю різноманітних публікацій, коментування, тощо) є аналіз та кластеризація текстової інформації користувачів для подальшої її класифікації за певними ознаками і визначення «портрету» користувача.

Методи ідентифікації користувачів у WEB-просторі

В цій статті на основі досліджень методів ідентифікації користувачів у WEB-просторі, піднімається проблема перевірка належності декількох аккаунтів одному і тому самому користувачу та ціль їх створення, через аналіз текстової інформації, що вводить користувач під час листування, коментування, та написання статей.

Ставиться задача розробки аналізатора текстової інформації користувача для подальшої класифікації та виявлення декількох аккаунтів одного і того самого користувача, що можуть нести певну загрозу.

Проблема ідентифікації та аутентифікації користувачів залишається відкритою, адже поки ще не існує єдиного методу для виявлення суб'єкта, що користується автоматизованою системою (АС) у поточний момент часу [3]. Звідси випливає актуальність в дослідженні даної теми та розробці нових методів ідентифікації.

Позитивним результатом аутентифікації є авторизація користувача, тобто надання йому прав доступу до ресурсів. Для доступу до різних АС можуть застосовуватися різні методи аутентифікації.

Залежно від міри довірчих стосунків, структури, особливостей мережі і віддаленості об'єкта перевірка даних наданих до АС може бути односторонньою або взаємною. Розрізняють однофакторну і строгу (двофакторну) аутентифікації. В однофакторних системах, найпоширенішими є парольні системи аутентифікації [4].

Основною метою використання методів для задач кластеризації та класифікації текстових документів, створення на їх основі засобів кластеризації та класифікації текстових даних користувача із застосуванням апарату теорії матриць, зокрема псевдообернення і проекціювання, функцій Ляпунова, та застосування їх до методів ідентифікації користувача у WEB-просторі є завдання класифікувати аккаунти користувача до класів певної спрямованості з високою точністю.

Розробка багатофакторної системи аналізу текстової інформації

Для виконання поставленої задачі пропонується до вже розробленої багатофакторної системи аналізу інформації та прийняття рішень [5] розробити та додати аналізатор текстів користувача, що буде кластеризувати тексти та по ним класифікувати користувачів до заздалегідь заданих класів загрози.

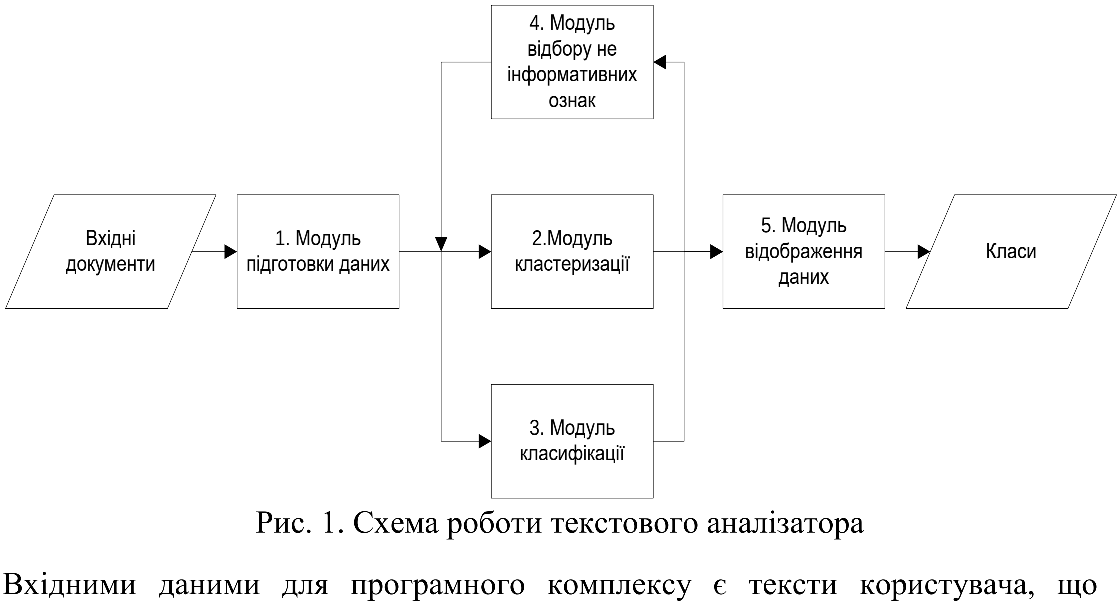


Рис. 1. Схема роботи текстового аналізатора

Вхідними даними для програмного комплексу є тексти користувача, що зберігаються у Mysql базі даних WEB-сайту з використанням кодування UNICODE.

Задачею попередньої обробки – є побудова вектору ознак. Основними засобами тут є видалення стоп-слів, приведення до основної словоформи.

В результаті будується вектор-ознак, засобами $ConfWeight$ [6], документу що складається зі всіх слів що входять в документ після застосування засобів попередньої обробки.

Виділимо основні етапи реалізації кластеризатора (метод K-гіперплощинної кластеризації):

1. Попередня обробка інформації;
2. Виділення характеристик: вибір властивостей, що характеризують об'єкти. Розрізняють кількісні характеристики і якісні. Основною характеристикою текстової інформації є слово й кількість його екземплярів у конкретному документі;
3. Визначення метрики. Метрика вибирається залежно від простору, де розташовані об'єкти, і неявних характеристик кластерів. Якщо всі координати об'єкта безперервні й речовинні, а кластери повинні являти собою щось подібне до гіперсфер, то використовується метрика Евкліда:

$$d_2(x_i, x_j) = (\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2)^{1/2} = \|x_i - x_j\|_2 \tag{1}$$

4. Розбиття об'єктів на групи: власне сама кластеризація;
5. Представлення результатів. Текстову інформацію, розбиту на підмножини зручніше представляти у вигляді папок (кластерів) – категорій і кластеризованих текстових документів усередині них.

На вхід модуля класифікації на етапі навчання подається матриця документів що представляє собою навчальну вибірку. Після навчання на вхід алгоритмів класифікації подається тестова вибірка у вигляді векторів-ознак документів.

Для вирішення завдань класифікації сигналів запропоновано новий підхід заснований на побудові спеціальних функцій Ляпунова за навчальною вибіркою і використанні природної метрики, за якою здійснюється вимірювання відстані до найближчого сусіда при розпізнаванні сигналів. Функції Ляпунова представляються у вигляді квадратичних форм у просторі ознак, у яких використовується матриця, що відображає розподіл точок навчальної вибірки за впорядкованими факторними напрямами.

Розглянемо в просторі ознак для досліджуваних сигналів чи деякого класу подій сукупність точок $x(j) \in R^m, j = \overline{1,n}$. Надалі при розгляді задач будемо використовувати відповідні проекційні операції, що визначаються наступними матрицями:

$$x(j) \in R^m, j = \overline{1,n} \tag{2}$$

$$R(\widetilde{X}^T) = \widetilde{X}^{+^T} - \widetilde{X}^{+} \tag{3}$$

$$\widetilde{X} = (x(1) - \hat{x} \mathrm{MLM} x(n) - \hat{x}), \hat{x} = \frac{1}{n}\sum_{j=1}^{n} x(j), \tag{4}$$

$\widetilde{X}^{+}$ – псевдообернена матриця до матриці $\widetilde{X}, I_m$ – одинична матриця в просторі R^m.

Тоді квадратичні форми мають наступний вигляд

$$(x - \hat{x})^T R(\widetilde{X}^T)(x - \hat{x}) \tag{5}$$

і відповідні їм еліпсоїдні контейнери – циліндри можуть бути представлені наступним чином

$$(x-\hat{x})^T R(\widetilde{X}^T)(x-\hat{x}) = c^2 \quad (6)$$

що мають деякі цікаві й важливі для застосування властивості.

На підставі сингулярного представлення матриці можна виділити наступне твердження, якщо

$$\widetilde{X} = \sum_{i=1}^{r} u_i v_i^T \lambda_i, r = rank\widetilde{X}, \widetilde{X}\widetilde{X}^T u_i = \lambda_i^2 u_i, \widetilde{X}^T \widetilde{X} v_i = \lambda_i^2 v_i, u_i^T u_j = \delta_{ij}, i, j = 1, r, \lambda_1^2 \geq ... \geq \lambda_r^2,$$

тоді мають місце співвідношення

$$(x-\hat{x})^T R(\widetilde{X}^T)(x-\hat{x}) = \sum_{i=1}^{r} \lambda_i^{-2}((x-\hat{x})^T u_i)^2, \lambda_i^2 = \sum_{j=1}^{n} (\widetilde{x}^T(j)u(j))^2, i = \overline{1,r} \quad (7)$$

$$\sum_{j=1}^{n} c_j^2 = r, c_j^2 = (x(j)-\hat{x})^T R(\widetilde{X}^T)(x(j)-\hat{x}), j = \overline{1,n}, \widetilde{x} = x - \hat{x} \quad (8)$$

Припустимо що $r = m$, тобто це означає, що в сукупності векторів $x(j), j = \overline{1,n}$ існує m лінійно незалежних між собою векторів.

Тоді для наступних еліпсоїдів

$$(x-\hat{x})^T R(\widetilde{X}^T)(x-\hat{x}) = c_j, j = \overline{1,n} \quad (9)$$

на яких розміщені точки $x(j), j = \overline{1,n}$, розглянемо еліпсоїд, що перебуває в середині сімейства еліпсоїдів

$$(x-\hat{x})^T R(\widetilde{X}^T)(x-\hat{x}) = c, c = \frac{1}{n}\sum_{j=1}^{n} c_j = \frac{m}{n} \quad (10)$$

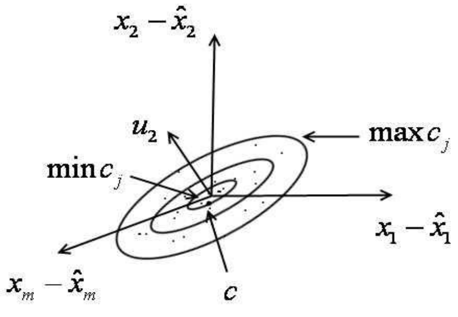


Рис. 2. Геометрична інтерпретація еліпсоїда

Цей еліпсоїд (2) можна описати еквівалентними рівняннями:

$$\frac{n}{m}(x-\hat{x})^T R(\widetilde{X}^T)(x-\hat{x}) = 1, \quad (11)$$

або $V(x,\hat{x},n,R)=1, V(x,\hat{x},n,R)=\dfrac{n}{m}(x-\hat{x})R(\widetilde{X}^T)(x-\hat{x}), R=R(\widetilde{X}^T)\in R^{m\times n}$.

Тут значення m як параметра не залежить від представників сукупності й тому зміни функцій Ляпунова $V(x,\hat{x},n,R)$, нижче досліджується по змінній x і відповідним параметрам $\hat{x},R,n$.

Якщо далі ми розглянемо дві множини точок зі своїми функціями Ляпунова [7] $V_1(x,\hat{x}(1),n(1),R_1)$ й $V_2(x,\hat{x}(2),n(2),R_2)$, то одержимо геометричну ілюстрацію приведену на рис. 5.2, яка відображає розміщення точок щодо поверхонь

$$V_1(x,\hat{x}(1),n(1),R_1)=1, \tag{12}$$

$$V_2(x,\hat{x}(2),n(2),R_2)=1. \tag{13}$$

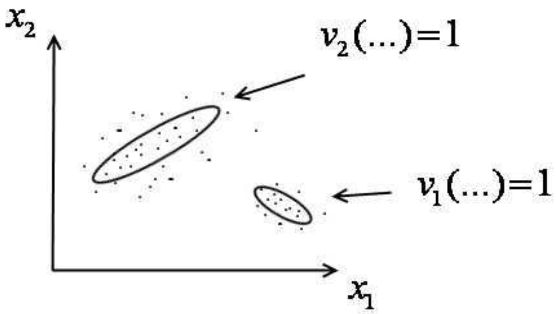


Рис. 3. Розміщення точок щодо поверхонь

Це означає, що відстані від точки x, щодо якої потрібно прийняти рішення якій із двох множин вона належить, до розглянутих центрів $\hat{x}(1)$, $\hat{x}(2)$ більш доцільно вимірювати не за допомогою Евклідової норми, а на підставі побудованих функцій Ляпунова.

При $V_1(x,\hat{x}(1),n(1),R_1)<V_2(x,\hat{x}(2),n(2),R_2)$ точка x відноситься до першої множини, і відповідно при протилежному знаку нерівності відноситься до другої множини.

У такий спосіб на етапі синтезу системи класифікації сигналів для l класів навчальна послідовність точок кожного k-го класу визначає функцію $V_k(k)=(x,\hat{x}(k),n(k),R_k)$ й процес розпізнавання далі здійснюється згідно зі схемою, представленої на рис. 4:

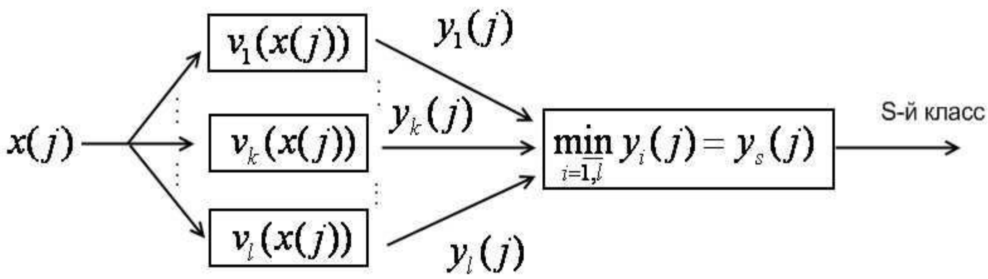


Рис. 4. Процес розпізнавання

Задачею тестування було перевірити чи належить вхідна текстова інформація до одного, чи декількох класів загрози, що були попередньо задані.

З кожної тематики було відібрано певну кількість документів в якості навчальної вибірки. Користувачі використовувались як тестові для визначення якості класифікації.

Загальний набір становив 543 документів. Помилка при визначенні конкретно кожного коментаря не визначалася, через складність цього підрахунку в ручну. Важливішим тут є підрахунок похибки при класифікації користувача.

Загальна кількість користувачів складала 30 аккаунтів. Загальна кількість унікальних входжень кожного аккаунта склала 51 профіль. З цих профілів вірно було класифіковано 50. Невірно – 11. Якщо брати загальну можливу кількість зв'язків між профілям та рубриками – похибка склала 17.73%.

Висновки

Для виконання поставленої задачі було розроблено автоматичний аналізатор текстової інформації користувачів на основі багатофакторної системи ідентифікації користувача у WEB-просторі.

Проведені тестування системи показали непоганий результат перевірки користувачів на відповідність заздалегідь запропонованим категоріям.

Розроблену систему можна широко використовувати на практиці, наприклад для виявлення інтернет «ботів» в соціальних мережах, формах та порталах новин.

Література

1. Галатенко В.А. Основы информационной безопасности. 2006. -С. 5-10.
2. Галатенко В.А. Информационная безопасность. [Електронний ресурс] // Открытые системы – 1996. – №4. – С. 79-86. – Режим доступу до журналу: http://www.osp.ru/data/www2/os/1996/04/40.htm.
3. Whitaker J.L., Bushman B.J. Online Dangers (2009): Keeping Children and Adolescents Safe. 66 Wash. & Lee Rev.1053
4. Куликова О.В. Методы и средства аутентификации в задачах обеспечения информационной безопасности корпоративных иформационных системах. 2009.
5. Марголін О.Г., Катеринич Л.О., Калашнікова А.А. Система прийняття рішень для ідентифікації користувачів //Вісник Київського університету. Серія: фіз. - мат. науки. Вип. 2, 2015. – С. 147-151.
6. Корлюк О.С. Методи з адаптацією параметрів моделей для класифікації текстової інформації: дис. канд. тех. наук : 01.05.02 / Корлюк Олександр Сергійович. – К., 2013. – 132 с.
7. Крак Ю.В., Кудін Г.І.Застосування методів розділення векторів ознак гіперплощиною в задачах розпізнавання елементів //Вісник Київського університету. Серія: фіз. - мат. Н.. Вип. 2, 2012. – С. 192-198.
8. Шатырко А.В. Качественный анализ систем регулирования нейтрального типа в условиях неопределенности с позиций функций Ляпунова// Доповіді НАНУ – 2012, №5. –С.43-48.

Literatura

1. Galatenko V.A. Osnovy informatsionnoy bezopasnosti. 2006. -С. 5-10.
2. Galatenko V.A. Informatsionnaya bezopasnost'. [Yelektronniy resurs] // Otkrytyye sistemy – 1996. – №4. – С. 79-86. – Rezhim dostupu do zhurnalu.: http://www.osp.ru/data/www2/os/1996/04/40.htm.
3. Whitaker J.L., Bushman B.J. Online Dangers (2009): Keeping Children and Adolescents Safe. 66 Wash. & Lee Rev.1053
4. Kulikova O.V. Metody i sredstva autentifikatsii v zadachakh obespecheniya informatsionnoy bezopasnosti korporativnykh iformatsionnykh sistemakh. 2009.
5. Marholin O.H., Katerynych L.O., Kalashnikova A.A. Systema pryynyattya rishen′ dlya identyfikatsiyi korystuvachiv //Visnyk Kyyivs′koho universytetu. Seriya: fiz. - mat. nauky. Vyp. 2, 2015. – S. 147-151.
6. Korlyuk O.S. Metody z adaptatsiyeyu parametriv modeley dlya klasyfikatsiyi tekstovoyi informatsiyi: dys. kand. tekh. nauk : 01.05.02 / Korlyuk Oleksandr Serhiyovych. – K., 2013. – 132 s.

7. Krak Yu.V., Kudin H.I. Zastosuvannya metodiv rozdilennya vektoriv oznak hiperploshchynoyu v zadachakh rozpiznavannya elementiv //Visnyk Kyyivsʹkoho universytetu. Seriya: fiz. - mat. nauky. Vyp. 2, 2012. – S. 192-198.
8. Shatyrko A.V. Kachestvennyy analiz sistem regulirovaniya neytral'nogo tipa v usloviyakh neopredelennosti s pozitsiy funktsiy Lyapunova// Dopovídí NANU – 2012, №5. –S.43-48.

RESUME

A.G. Margolin

Information analysis system for user text messages

This article is based on methods and tools for identifying subjects of an automated system, through the analysis of textual information that the user enters in the correspondence, commentary, and writing articles, research.

The problem of the user information text analyzer development for further classification and identification of few accounts that belong to the same user, that may be a threat, is posed.

This problem remains unresolved, because there is no single method for identifying a person who is using the computer at the moment. This implies the relevance of the current topic research and development of new methods of identification.

For solving this problem, the development of a user text analyser and adding it to the multifactor decision-making system, which was developed before, is proposed. The text analyzer will cluster texts and classify users into predefined classes of danger.

To build a system of methods used for clustering and classification tasks of text documents

Methods for text documents clustering and classification problems are used for building the system. On this basis, tools for user text data clustering and classification were used, with the use of the matrix theory, particularly the Lyapunov functions projection and pseudoinversion. It has been applied to the methods of user identification in the WEB.

As a result, an automatic user messages text analyser has been developed, based on the multifactor system for identifying user in the WEB, testing was performed.

The task of testing was to check whether the input text information refers to one or more classes of threats that have been predefined.

A number of documents from each subject were selected as a training set. User accounts used as a test to determine the quality classification.

The total set of documents was 543. The error in determining each comment specifically has not been determined due to the difficulty of counting. More important is the error in the user classification calculation.

The total number of users was 30 accounts. The total number of unique occurrences of each account was 51 profile. These profiles were classified correctly - 50. Incorrect - 11. The total possible number of connections between the profiles and sections error was 17.73%.

Testing of the system showed a good result of verifying users for compliance with predefined categories.

The system can be used in practice, for example to identify Internet "bots" in social networks, forums or news portals.

Надійшла до редакції 29.11.2016

